Visual Entity Linking via Multi-modal Learning (Postprint)

Authors: Qiushuo Zheng, Hao, Wen, Meng, Wang, Guilin, Qi, Guilin, Qi

Date: 2022-11-28T00:00:00+00:00

Abstract

Existing visual scene understanding methods mainly focus on identifying coarse-grained concepts about the visual objects and their relationships, largely neglecting fine-grained scene understanding. In fact, many data-driven applications on the Web (e.g., news-reading and e-shopping) require accurate recognition of much less coarse concepts as entities and proper linking them to a knowledge graph (KG), which can take their performance to the next level. In light of this, in this paper, we identify a new research task: visual entity linking for fine-grained scene understanding. To accomplish the task, we first extract features of candidate entities from different modalities, i.e., visual features, textual features, and KG features. Then, we design a deep modal-attention neural network-based learning-to-rank method which aggregates all features and maps visual objects to the entities in KG. Extensive experimental results on the newly constructed dataset show that our proposed method is effective as it significantly improves the accuracy performance from 66.46% to 83.16% compared with baselines.

Full Text

Preamble

RESEARCH PAPER

Visual Entity Linking via Multi-modal Learning

Qiushuo Zheng[1], Hao Wen[2], Meng Wang[2],[3] & Guilin Qi[2],[3]†

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

[2]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China

[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China

Keywords: Knowledge graph; Multi-modal learning; Entity linking; Learning to rank; Knowledge graph representation

Citation: Zheng, Q.S., et al.: Visual entity linking via multi-modal learning. Data Intelligence 4(1), 1-19 (2022). doi: 10.1162/dint_a_{00114}
Received: September 15, 2021; **Revised:** November 1, 2021; **Accepted:** November 18, 2021

Abstract

Existing visual scene understanding methods mainly focus on identifying coarse-grained concepts about visual objects and their relationships, largely neglecting fine-grained scene understanding. In fact, many data-driven applications on the Web (e.g., news-reading and e-shopping) require accurate recognition of much less coarse concepts as entities and proper linking them to a knowledge graph (KG), which can take their performance to the next level. In light of this, we identify a new research task: visual entity linking for fine-grained scene understanding. To accomplish this task, we first extract features of candidate entities from different modalities, i.e., visual features, textual features, and KG features. Then, we design a deep modal-attention neural network-based learning-to-rank method which aggregates all features and maps visual objects to entities in KG. Extensive experimental results on a newly constructed dataset show that our proposed method is effective as it significantly improves accuracy performance from 66.46% to 83.16% compared with baselines.

1. Introduction

Visual scene understanding is inevitably regarded as one of the core functions of next-generation machine intelligence, and it has been evolving to meet increasing demands not limited to object detection from images and video clips, but more often than not, for grasping the story behind the pixels. Due to the representation power of graph structure, the scene graph [1] has been proposed to harness external knowledge for better scene understanding. Thus parsing an image into a scene graph has attracted much research attention over the past few years. Although great efforts have been made on generating scene graphs from images, existing methods only detect visual objects at a coarse-grained concept level (i.e., categories), which sometimes offers little help for deeper scene understanding. In many practical scenarios, such as news-reading and e-shopping, we require entity-level visual object detection for subsequent question answering or recommendation systems.

Motivating example: Consider the following two scenarios: (1) An online user is reading sports news about basketball and wants to distinguish Yao Ming and Tracy McGrady in a group photo, as shown in Figure 1 [Figure 1: see original paper]. However, even world-leading object recognition systems cannot guarantee to provide the correct answer. (2) Another user is interested in Tracy McGrady' s shoes and would like to know the specific signature sneaker, but

existing image search engines like Bing.com can only recognize them as white shoes. To accomplish these user tasks, we need more detailed auxiliary information to complement visual learning. This complementary information can be obtained from comprehensive multi-modal knowledge graphs (KG), such as DBpedia and IMGpedia. If entities in KG are successfully linked to objects in the image, we can answer the question with the correct name (i.e., Tracy McGrady) in Case 1 and precisely recommend to the user the specific shoe brand (i.e., Adidas T-MAC 4) in Case 2.

Challenges: To achieve fine-grained visual entity linking in scene understanding, several challenges exist: (1) It is difficult to recognize all visual objects from an image solely based on visual information. Even state-of-the-art fine-grained object detection models trained with enormous labeled samples cannot guarantee accurate classification of all fine-grained classes. (2) When linking entities extracted from KGs to visual objects detected from an image, the visual entity linking algorithm needs to effectively exploit heterogeneous features and utilize cross-modal correlations to achieve disambiguation and find the accurate entity in KG for each visual object in an image.

Solutions: In this paper, we propose a novel framework to achieve visual entity linking in visual scene understanding. Specifically, we first generate a coarse-grained scene graph for an image and extract visual features of objects by employing a VGG-16 network. Then, we utilize the Gated Recurrent Unit (GRU) language method to extract textual features of objects from image captions and discover candidate KG entities through named mention matching. After extracting KG features for candidate entities, we propose a deep modal-attention neural network-based learning-to-rank method to aggregate all features and map visual objects to entities in KG.

Contributions: The contributions of this paper are summarized as follows:

- We are the first to consider visual entity linking in scene graphs and have constructed a new large-scale dataset for this challenging task.
- We propose a novel framework to learn features from three different modalities and simultaneously design a learning-to-rank based visual entity linking model, which aggregates different features through a deep modal-attention neural network.
- We conduct extensive experiments to evaluate our visual entity linking model against state-of-the-art methods. Results on the constructed dataset show that our proposed method is effective as it significantly improves accuracy performance from 66.46% to 83.16% compared with baselines.

2. Related Work

This section discusses existing related research from the following aspects: entity linking, visual scene understanding, and multi-modal learning.

2.1 Entity Linking

Entity linking is the task of mapping mentions in text to corresponding entities in KG. Conventional models are usually distinguished by the supervision they require, i.e., supervised or unsupervised methods. Supervised methods [2, 3, 4, 5] utilize annotated data to train binary classifiers or ranking models to realize entity disambiguation. Unsupervised methods [6, 7, 8, 9, 10] generally use similarity measures between mentions in text and entities in KG.

In contrast to traditional research in the textual domain, the visual entity linking task has the following differences: (1) The form of visual representation is much more complicated than forms that may appear in textual content, and (2) Different modalities have different characteristics.

2.2 Visual Scene Understanding

Visual scene understanding includes many kinds of work, such as traditional computer vision tasks like image recognition [11, 12] and higher-level scene reasoning tasks like scene graph generation. In recent years, considerable progress has been made on many sub-issues of the overall visual scene understanding problem. Since early work [1, 13, 14] generated visually-grounded graphs over objects with their relationships in an image, many models have been proposed to improve performance, such as adding prior probability distribution [13, 15, 16] and introducing message passing mechanisms [17, 18, 19].

2.3 Multi-modal Learning

Multi-modal learning [20, 21, 22] focuses on learning with contextual information from multiple modalities in a joint model. Recent relevant tasks to our work include cross-modal entity disambiguation and entity-aware captioning. Ref. [23] built a deep zero-shot multi-modal network for social media post disambiguation, but they remain limited to linking textual entities in posts to the knowledge base. Refs. [24, 25, 26, 27, 28] proposed multi-modal entity-aware models to achieve image caption generation, which is also different from our visual entity linking task.

Recently, Li et al. [29] presented a multimedia knowledge extraction system that enabled graph query search and retrieved multimedia evidence. However, when dealing with visual entity linking, it adopts redundant rules for different entity types, increasing model complexity.

Entity linking is an important branch in natural language processing, but existing models cannot solve the problem of multi-modal learning in entity linking. This paper proposes a joint model to provide ideas for solving this problem.

3.1 Problem Formulation

In this section, we describe our multi-modal learning model for visual entity linking in detail. As illustrated in Figure 2 [Figure 2: see original paper], the proposed model includes the feature extraction module and the visual entity linking module.

For the training process, given a dataset of input samples for the visual entity linking task, the number of input samples is m, denoted by $\{(x_i, t_i, y_i)\}_{i=1}^{m}$, where x_i is the combination of the i-th image' s visual features and textual features, t_i represents the j-th bounding box of the i-th image, and y_i is the total set of ground truth entities. Each input sample x_i has corresponding textual information t_i and length L_t, and entity y_i in the multi-modal knowledge graph KG linked to x_i. We aim to learn a transformed function $f(\cdot)$ that satisfies:

$$y = \arg\max \mathrm{sim}(f(x), y)$$

where $f(\cdot)$ is a transformed function that projects input samples x_v and x_t into the same space as y, and $\mathrm{sim}(\cdot)$ generates a similarity score between prediction and ground truth.

For the testing process, given a test sample s_p including the image with corresponding caption, i.e., $s_t = (v_p, t_p)$. The bounding boxes of the test image are generated by the scene graph, represented by v_p^q for the q-th bounding box of the p-th image. Then, the visual entity linking task is to match the bounding box v_p^q to the entity y_p^q in KG by the function.

3.2 Feature Extraction Module

The feature extraction module aims to extract features from three modalities as follows:

Visual features: To link image bounding boxes x_v to the given KG, we first used the outperforming method from [30] for generating bounding boxes in the scene graph. Then, we used the VGG-16 network [31] for visual feature extraction of image bounding boxes. The final layer representation $e(x_v)$ of the VGG-16 network is transformed into low dimensions, which describes features of an image bounding box.

Textual features: To extract textual features from image caption x_t, we encoded the caption by a GRU language model [32] with distributed word semantics embeddings $e(x_t)$. We used the following implementation for the GRU:

$$z_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1})$$
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$
$$\tilde{h}_t = \tanh(W_{xh}x_t + r_t \odot (W_{hh}h_{t-1}))$$

$$e(x_t) = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

where h_t is a hidden layer output at decoding step t, r_t controls the influence of the hidden layer unit h_{t-1} at the previous moment on the current word x_t, z_t decides whether to ignore the current word x_t, and $e(x_t)$ is the output textual vector representation of GRU at the last decoding step $t = T$.

Because pre-trained models can effectively represent semantic distribution of words in a sentence, we used pre-trained embeddings from the GloVe model [33] in the GRU sentence encoder.

Similar to traditional entity linking models, we also need to obtain the list of named entities in the image caption x_t. We implemented a Named Entity Recognizer (NER) based on BERT [34], Bi-LSTM, and Conditional Random Fields (CRF), as shown in Figure 3 [Figure 3: see original paper]. The Bi-LSTM extracts higher-level structural information, and CRF is used as a sequence classifier. We fine-tuned the model on our dataset to achieve the best recognition results. Once we established the named entity list with NER, we sent the list to a SPARQL query engine and obtained candidate entities in KG.

KG features: To generate linked candidate entities, we proposed a matching algorithm based on rules. Then, we obtained each candidate entity' s image embedding $e(y_v)$ and structural text information embedding $e(y_t)$ as its KG features.

Because of the inaccuracy and incompleteness of entity mentions in caption x_t (i.e., abbreviations and nicknames for person names), directly sending entity mentions from captions to the KG SPARQL query engine may not establish a complete candidate entity list. To address this, we implemented a rule-based candidate entity list generator using partial matching strategy and four rules: (1) The entity name shares several words with the entity mention; (2) The entity name is wholly contained in or contains the entity mention; (3) The entity name exactly matches the first letters of all words in the entity mention; and (4) The entity name has high string similarity over 80% with the entity mention in Levenshtein distance.

For KG visual embedding $e(y_v)$ of y, we also used the same VGG-16 network to extract dense visual features. For KG structural text embedding $e(y_t)$ of y, we used DBpedia [35] as our multi-modal knowledge graph to obtain embedding with the complex model proposed by [36], which is state-of-the-art. The KG structural text embeddings are learned by the following score function to measure a fact $\langle h, r, t \rangle$ in KG:

$$f(h, r, t) = \mathrm{Re}(\mathbf{h}\mathrm{diag}(\mathbf{r})\bar{\mathbf{t}})$$

where $\bar{\mathbf{t}}$ is the conjugate of $\mathbf{t}$ and $\mathrm{Re}(\cdot)$ means taking the real part of a complex value. $\mathbf{h}$ and $\mathbf{t}$ are entities in KG and may be possible matched entities for y.

3.3 Visual Entity Linking Module

In this module, we aggregated all three modality features to predict the best matched KG entity for each image bounding box. We proposed a supervised visual entity linking module using visual confidence and textual confidence.

The confidence of visual similarity for the i-th bounding box and j-th candidate entity is calculated between the image feature vector $e_v(x)$ and visual feature $e_v(y)$ extracted from KG. The confidence of textual similarity for the i-th bounding box and j-th candidate entity is calculated between the sentence vector $e_t(x)$ and structural text information embedding $e_t(y)$.

The loss function consists of two parts, where L_t is the supervised max-margin ranking loss for KG entity prediction on textual features, and L_v is the max-margin ranking loss on visual features. l_t and l_v denote hyper-parameters to tune the function. conf_v is the confidence score in the visual modality, and conf_t is the confidence score in the textual modality. Through our max-margin ranking loss function, the confidence of the correct linking entity $\text{conf}(y)$ should be higher than that of any other candidate entity $\text{conf}(y')$ with margin c, where $[x]_+$ denotes the positive part of x. m is the number of samples, and s is the serial number of the input sample. $f(\cdot)$ is a function that projects textual embeddings into KG structural embedding space. $\text{conf}_t(y_i)$ is the confidence score between caption textual embedding $e(x_t)$ and KG structural text embedding of i-th candidate entity in textual modality, and $\text{conf}_v(y_i)$ is the confidence score between image visual embedding $e(x_v)$ and KG visual embedding of i-th candidate entity $e_v(y)$ in visual modality.

To learn different weights of modalities, we formulated the modal-attention module as follows, which selectively weakens or magnifies different modalities:

$$\mathbf{r} = \text{softmax}(\mathbf{W}_a \mathbf{x})$$

$$\mathbf{x}_{\text{final}} = \sum_{m \in \{t,v\}} a_m \mathbf{x}_m$$

where $\mathbf{r}$ is an attention vector, and $\mathbf{x}_{\text{final}}$ is the final context vector that reasonably focuses on different modalities.

At test time, the following entity-predicting nearest neighbor (1-NN) classifier is used for prediction, where l_1 and l_2 are hyper-parameters:

$$\hat{y} = \arg\max_{y \in \text{KG}} \text{sim}(f(x), y)$$

4. Experiments

We constructed a new dataset for the task of visual entity linking and compared our model with state-of-the-art methods on this dataset.

4.1 Experiments Setting

Datasets. For the visual entity linking task, we need to link image bounding boxes to specific entities in KG, which goes beyond identifying object categories. However, most existing computer vision datasets contain no named entities in images or captions. Therefore, we built a new dataset, namely Visual Entity Linking Dataset (VELD), composed of 39,000 news image and textual caption pairs with links to KG entities, manually labeled by expert human annotators (entity types: PER, LOC, ORG). In total, we gathered 39,000 images with textual captions, randomly split into 31,000 for training, 4,000 for validation, and 4,000 for testing.

It is meaningless for the visual entity linking task if no words in captions describe named entities. Therefore, we performed a filtering program on named entities in VELD to remove news data that does not contain named entities, ensuring that named entities appear in sentences at 100%. Key aspects are summarized in Table 1 . The VELD dataset, similar to BreakingNews [37], exhibits longer average caption lengths than image-caption datasets like MSCOCO [38], indicating that news captions tend to be more descriptive.

Tasks. Given an image bounding box and an accompanying caption, our goal is to link the image bounding box to corresponding KG entities in DBpedia 2016 by utilizing visual features, textual features, and KG features.

Evaluation metrics. The primary metric of our evaluation is the accuracy of visual entity linking to KG entities. Accuracy is defined as in Equation (17):

$$\text{accuracy} = \frac{\text{correctly linked entity mentions}}{\text{all links generated by our method}}$$

Implementation details. We initialized the NER stream of our model with a BERT language model pre-trained on English Wikipedia. Specifically, we used the BERTBASE model which has 12 layers of transformer blocks with each block having a hidden state size of 768 and 12 multi-head attentions. We trained on four 2080Ti GPUs with a total batch size of 256 for 20 epochs. We used Adam optimizer with an initial learning rate of 0.001. We employed a decay learning rate schedule with warm-up to train the model.

Parameters. We used the following search spaces to adjust parameters of each modality (bold indicates final model choices): VGG-16 embedding dimensions: {128, 256, **512**, 1024}, GRU hidden states: {50, 100, 128, **150**, 200}, KG image embedding dimensions: {128, 256, **512**, 1024}, x-dimensions: {50, 100, **150**, 200, 250}, l_1: {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, **0.9**} and l_i: {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, **0.9**}. We optimized parameters by Adam with batch size = 10, learning rate = 0.001, $b_1 = 0.9$, $b_2 = 0.999$ and $\epsilon = 10^{-8}$.

4.2 Baselines

Because our proposed task is relatively novel, related models available for comparison are especially limited. We report performance of the following state-of-the-art entity linking and visual object recognition methods as baselines, as well as several configurations of our proposed method to examine contributions of each component (T: textual, V: visual, and KG: knowledge graph).

- **Huawei API** and **Tencent API** (V and KG) use deep neural network models to recognize people.
- **CoAtt [39]** (T and KG) uses a type-aware co-attention model for entity disambiguation.
- **Falcon [40]** (T and KG) performs joint entity linking of short text by leveraging several fundamental principles of English morphology.
- **CDTE [41]** (T and KG) proposes a neural, modular entity linking method using multiple sources of information for entity linking.
- **GENRE [42]** (T and KG) system realizes entity retrieval by generating entity names in a token-by-token auto-regressive manner from left to right, with results affected by context.
- **DZMNED [23]** (T, V and KG) uses an attention LSTM model for multimodal NED task in social media posts.
- **(Proposed) Our method** (T, V and KG) is the proposed method as described in Figure 2.
- **(Proposed) Our method without visual features** (T and KG) only uses textual features extracted from captions and KG.
- **(Proposed) Our method without textual features** (V and KG) only uses visual features extracted from images and KG.
- **(Proposed) Our method with a smaller sized KG** (T, V and KG)

4.3 Results

Comparison patterns. First, because of the novelty of visual entity linking, we need to rely on existing models to build comparative experimental methods. Next, we explain experimental settings and how to achieve relative fairness through settings under some unbalanced experimental conditions, such as V+KG modalities, V+T modalities, and T+KG modalities.

Intuitively, in V+KG modalities, the visual training data of their recognizer network performs the same work as KG. These massive image data and image resources in KG modality are equivalent and have the same effect. In the first two experiments of Table 2 , we used V modality to replace V+KG modalities to realize corresponding experiments.

V+T modalities cannot output results defined in the task because of lacking KG entity links. Ignoring the KG modality and target entities, the entity linking task cannot continue, so it cannot be used as a comparative experiment. Therefore, in our experiments, due to lack of target entities, we did not choose corresponding V+T modalities for comparative analysis.

Short-text entity linking based on KG (T+KG modalities) cannot link KG entities to corresponding image bounding boxes. For comparison, we first used scene graph method to generate corresponding bounding boxes, then randomly connected entities in the candidate list to entity bounding boxes. Simultaneously, we multiplied each accuracy rate by the number of candidate entities per entity bounding box to ensure fairness of accuracy rate. By multiplying accuracy of visual entity linking in T+KG modalities by the number of candidate entities, we eliminated error caused by random connection of candidate entities in T+KG modality experiments.

Main results. Table 2 shows Top-1, 3, 5, and 10 candidate entity list retrieval accuracy results on VELD dataset. The first two experiments use information from visual modality and knowledge graph modality. Through experimental results, we proved that existing deep neural networks based on static offline training cannot complete visual entity linking well. Due to limitations of training dataset, it is difficult to build a dataset containing image resources of all entities in the open domain, which proves our model' s validity from another perspective.

The third to fifth experiments are based on features from textual modality and knowledge graph modality for visual entity linking, and through a series of post-processing, linking of target frames is not affected by visual features. Experimental results show there is still a large gap between textual modality and our full model.

Compared with simple visual object recognition methods and textual entity linking methods using text and KG as support, we found that our proposed method significantly outperforms these baselines. The reason is that we jointly fused three kinds of features from different modalities rather than simple modality-based linking. Another convincing point is that by applying similar multi-modal learning model DZMNED on VELD dataset, results show they only achieved 66.46% on Top-1 accuracy measure, while our model reached 83.16%, demonstrating our model' s great advantage in visual entity linking task.

Visualization of modality attention. Figure 4 [Figure 4: see original paper] visualizes the modality attention module of our model, where we list each entity (each column) of some test samples, with amplified modality represented by darker color and attenuated modality shown by lighter color. We intuitively analyze from experimental results that more relevant modalities in visual entity linking will be emphasized through the modality attention module. Specifically, we used alignments between different modalities from VELD dataset test set.

For our multi-modal visual entity linking model (V+T+KG modalities), we confirm from experimental results that the modality attention module has successfully enhanced function of relevant modal information (e.g., in similar celebrity entity linking) and amplified relevant modality-based contexts in prediction.

In the first row example of Figure 4, "Jobs, Apple's founder, attended the launch of the new iPhone" , we first generated candidate entity list of "Jobs" , "Apple" , "iPhone" . For entity linking of "Jobs" entity, we learn that influence of visual

modality is higher than textual modality according to color depth of modalities. For other two entities "Apple" and "iPhone" , influence of visual modality is much lower than textual modality. Because there are few candidate entities for "Apple" and "iPhone" , relying only on textual modality we can easily find KG entity corresponding to contextual semantics, but there are many related entities for "Jobs" entity, so we need to use feature vector of visual modality for entity linking task, which is why different entity categories have different modality weights.

In the second example, "Curry won the NBA Championship for the Golden State Warriors at Auckland Stadium" , the candidate entity list comprises "Curry" , "Golden State Warriors" and "NBA" . For "Curry" entity, we found that modality attention module mainly concentrates on visual modality information. For "Golden State Warriors" entity, proportion of visual modality information and textual modality information is roughly equal, while for "NBA" entity, it mainly depends on textual modality signals.

In the third case, for person category like "Francis" , modality attention successfully focuses on visual modality and attenuates distracting signals, and for "Oscar" , visual modality information and textual modality information have equal status.

Ablation study. To evaluate effectiveness of different modules, we conducted several ablation experiments shown in Table 3 . We validated effects of features in three modalities: visual, textual, and KG.

Because our experimental result is a comprehensive expression of multiple modalities (i.e., basis of visual entity linking is image bounding box, and caption description generates candidate entity list), we used KG entities for links. Therefore, our input data is unchanged, and only parameters of corresponding part (l_1 and l_2) are adjusted to zero in confidence calculation to achieve elimination of particular modality feature. For smaller knowledge graph, we chose KG embeddings learned from 1M KG subset. Corresponding experiment results are shown in Table 3.

From experimental results, we find that features of each modality contribute certain amount to visual entity linking performance. Lack of any modality feature significantly reduces performance in terms of accuracy metric. Lack of visual features reduces Top-1 accuracy to 56%. Absence of text features decreases Top-1 accuracy to 73%. Reducing KG scale decreases Top-1 accuracy to nearly 60%. These results suggest that jointly utilizing multi-modal features obtains best experimental linking results.

Error analysis. In example "Robert Downey Jr. plays Iron Man in the movie" , our model links image bounding boxes to actor Robert Downey Jr., while ground-truth links to Iron Man. This means our model sometimes outputs erroneous results in situations where one person has multiple roles in KG.

Therefore, in some instances, we need to set rules and constraints to obtain bet-

ter results. In other scenarios, when there is occlusion or concealment in image, deviations occur. For example, in another case, we can easily link Suárez in KG to image bounding box, but it is much more difficult to link Messi because of visual occlusion in image. Such error reasons can be attributed to incompleteness of image information, and from another aspect, also indicates importance of image features for visual entity linking.

5. Conclusion and Future Work

In this paper, we introduced a new task called visual entity linking, which links KG entities to corresponding image bounding boxes, and addressed the problem through a novel framework. The proposed framework first extracts features from three modalities (visual, textual, and KG). Then, a deep modal-attention neural network is employed for linking entities to corresponding image bounding boxes.

We constructed a new dataset VELD for visual entity linking experiments. Experimental results show that our model achieved state-of-the-art results. Moreover, through extensive ablation experiments, we demonstrated efficacy of our method.

In future work, a possible improvement direction is to utilize structural embedding features of scene graph to improve visual entity linking performance. In addition, we hope our model becomes a generic framework for visual entity linking task, but constructing an ideal complete KG including all entities in the world is impossible. Therefore, requirements and effects of visual entity linking need to be determined according to specific applications.

References

[1] Johnson, J., et al.: Image retrieval using scene graphs. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 3668-3678 (2015)

[2] Shen, W., et al.: Linden: Linking named entities with knowledge base via semantic knowledge. In: International World Wide Web Conference, pp. 449-458 (2012)

[3] Lin, T., et al.: Entity linking at Web scale. In: The Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, pp. 84-88 (2012)

[4] Nguyen, T.H., et al.: Joint learning of local and global features for entity linking via neural networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2310-2320 (2016)

[5] Chen, Z., Ji, H.: Collaborative ranking: A case study on entity linking. In: Empirical Methods in Natural Language Processing, pp. 771-781 (2011)

[6] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia

data. In: Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716 (2007)

[7] Pan, X., et al.: Unsupervised entity linking with abstract meaning representation. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1130–1139 (2015)

[8] Chisholm, A., Hachey, B.: Entity disambiguation with Web links. Transactions of the Association for Computational Linguistics 3, 145–156 (2015)

[9] He, Z., et al.: Efficient collective entity linking with stacking. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 426–435 (2013)

[10] Cheng, X., Roth, D.: Relational inference for Wikification. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1787–1796 (2013)

[11] He, K., et al.: Deep residual learning for image recognition. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 770–778 (2016)

[12] Gorniak, P., Roy, D.: Grounded semantic composition for visual scenes. Journal of Artificial Intelligence Research 21, 429–470 (2004)

[13] Lu, C., et al.: Visual relationship detection with language priors. In: European Conference on Computer Vision, pp. 852–869 (2016)

[14] Yao, B., Li, F.-F.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17–24 (2010)

[15] Farhadi, A., et al.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29 (2010)

[16] Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

[17] Xu, D., et al.: Scene graph generation by iterative message passing. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 5410–5419 (2017)

[18] Ramanathan, V., et al.: Learning semantic relationships for better action retrieval in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1100–1109 (2015)

[19] Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1745–1752 (2011)

[20] Baltrusaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine

Intelligence 41(2), 423–443 (2018)

[21] Cao, N.D., et al.: Autoregressive entity retrieval. In: The 9th International Conference on Learning Representations (ICLR 2021). Available at: https://openreview.net/forum?id=5k8F6UU39V/. Accessed 11 December 2021

[22] Zhu, Y., et al.: Knowledge perceived multi-modal pretraining in e-commerce. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2744–2752 (2021)

[23] Moon, S., Neves, L., Carvalho, V.: Multimodal named entity disambiguation for noisy social media posts. In: Annual Meeting of the Association for Computational Linguistics, pp. 2000–2008 (2018)

[24] Biten, A.F., et al.: Good news, everyone! context driven entity-aware captioning for news images. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 12466–12475 (2019)

[25] Lu, D., et al.: Entity-aware image caption generation. arXiv preprint arXiv:1804.07889 (2018)

[26] Antol, S., et al.: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)

[27] Tsai, Y.-H.H., et al.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference Association for Computational Linguistics, pp. 6558–6569 (2019)

[28] Sil, A., et al.: Multi-lingual entity discovery and linking. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 22–29 (2018)

[29] Li, M., et al.: Gaia: A finegrained multimedia knowledge extraction system. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 77–86 (2020)

[30] Zhang, J., et al.: Graphical contrastive losses for scene graph parsing. In: Internaltional Conference on Computer Vision and Pattern Recogintion, pp. 11535–11543 (2019)

[31] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[32] Chung, J., et al.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)

[33] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[34] Devlin, J., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[35] Auer, S., et al.: DBpedia: A nucleus for a Web of open data. In: International Semantic Web Conference & Asian Semantic Web Conference, pp. 722–735 (2007)

[36] Trouillon, T., et al.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning, pp. 2071–2080 (2016)

[37] Ramisa, A., et al.: Article annotation by image and text processing. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(5), 1072–1085(2017)

[38] Lin, T.-Y., et al.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014)

[39] Nie, F., et al.: Mention and entity description co-attention for entity disambiguation. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018), pp. 5908–5915 (2018)

[40] Sakor, A., et al.: Old is gold: Linguistic driven approach for entity and relation linking of short text. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2336–2346 (2019)

[41] Gupta, N., Singh, S., Roth, D.: Entity linking via joint encoding of types, descriptions, and context. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2681–2690 (2017)

[42] De Cao, N., et al.: Autoregressive entity retrieval. arXiv preprint arXiv:2010.00904 (2021)

Author Biographies

Qiushuo Zheng is a graduate student at School of Cyber Science and Engineering, Southeast University. He received a bachelor' s degree from Southeast University. His main research interests are multi-modal learning and downstream applications of knowledge graph. ORCID: 0000-0003-4921-2218

Hao Wen is an undergraduate student at the School of Computer Science and Engineering, Southeast University. Currently, his research interests mainly include information retrieval, entity linking and multi-media research. ORCID: 0000-0003-3165-3859

Meng Wang is an Assistant Professor in the Knowledge Graph & AI Research Group, School of Computer Science and Engineering, Southeast University (SEU), China. He is also a SEU Zhishan Young Scholar. He obtained his doctoral degree from the Department of Computer Science and Technology, Xi' an Jiaotong University in 2018, under supervision of Prof. Jun Liu. He was a visiting scholar, working with Prof. Xue Li and Prof. Xiaofang Zhou, in the DKE Lab at University of Queensland, Australia in 2016. His research area is in knowledge graph, semantic search, natural language processing (NLP), and cross-modal data. ORCID: 0000-0002-2293-1709

Guilin Qi is a Professor at Southeast University, China, where he also serves as Director of the Institute of Cognitive Intelligence of Southeast University. He was supported by the Six Talent Peak Programs of Jiangsu Province. At present, he is Deputy Director of the Language and Knowledge Computing Professional Committee of Chinese Information Society and Deputy Director of the Knowledge Organization Professional Committee of China Science and Technology Information Society. He was a visiting professor at Griffith University in Australia and a visiting professor at the First University of Toulouse in France. He graduated from Yichun University, majoring in mathematics, in 1998, obtained a Master' s degree from the Mathematics and Information Department of Jiangxi Normal University in 2002 and a Doctor' s degree in Computer Science from Queen' s University of Belfast in 2006. ORCID: 0000-0002-1957-6961

Note: Figure translations are in progress. See original paper for figures.